How to exploit paralinguistic features to identify acronyms in texts?

Mathieu Roche

UMR TETIS, Cirad, Irstea, AgroParisTech,
500, rue J.F. Breton, 34093 Montpellier Cedex 5, France
Mathieu.Roche@cirad.fr

LIRMM, CNRS, Univ. Montpellier 2,
161, rue Ada, 34000 Montpellier, France
Mathieu.Roche@lirmm.fr

Abstract

This paper addresses the issue of acronym dictionary building. The first step of the process identifies acronym/definition candidates, the second one selects candidates based on a letter alignment method. This approach has two advantages because it enables (1) to annotate documents, (2) to build specific dictionaries. More precisely, this paper discusses the use of a specific linguistic concept, *the gloss*, in order to identify candidates. The proposed method based on paralinguistic markers is independent of languages.

Keywords: text mining, acronym expansion

1. Introduction

Acronyms are numerous in specialized domain, e.g. biomedical and agronomy documents (Chang et al., 2002). An acronym is a set of characters corresponding to the first letters of a group of words, for instance, the acronym *FAO* is associated with the definition *Food and Agriculture Organization*. This paper summarizes a method to identify acronyms and expansions in documents. This automatic recognition enables to annotate these elements in texts.
This work deals with the use of paralinguistic features in order to identify acronym/definition couples.

After the description of related work in the following section, Section 3. describes our approach based on 2 steps: Extraction of acronym/expansion candidates (Section 3.1.), Filtering of candidates (Section 3.2.). Finally, before Discussion and Conclusion sections, experiments of our approach are detailed in Section 4.

2. Related work

Among the several existing methods for acronym extraction in the literature, some significant work need to be mentioned. The acronym detection involves recognizing a character chain as an acronym and not as an unknown or misspelled word. Most acronym detecting methods rely on using specific linguistic markers.

Yates' method (Yeates, 1999) involves the following steps: First, separating sentences by segments using specific markers (brackets, points) as frontiers. Then the acronym/expansion couples are tested. The acronym/definition candidates are accepted if the acronym characters correspond to the first letters of the potential definitions words. The last step uses specific heuristics to select the relevant candidates. These heuristics rely on the fact that acronyms length is smaller than their expansion length, that they appear in upper case, and that long expansions of acronyms tend to use stop-words such as determiners, prepositions, suffixes and so forth. Therefore, the pair "FAO/Food and Agriculture Organization" is valid according to these heuristics.

Other studies (Chang et al., 2002; Larkey et al., 2000) use similar methods based on the presence of markers associated with linguistic and/or statistical heuristics. In this context (Okazaki and Ananiadou, 2006) propose statistical measurements from the terminology extraction area. Okazaki and Ananiadou apply the C-value measure (Frantzi et al., 2000; Nenadic et al., 2003) initially used to extract terminology. It favors a candidate term that doesn't appear often in a longer term. For instance, in a specialized corpus (i.e. Ophthalmology), the authors discovered that the term "soft contact" was irrelevant, while the frequent and longer term "soft contact lens" is relevant. An advantage of C-value measure is its independence from characters alignment (actually, a lot of acronyms/definitions are relevant while the letters are in a different order, e.g. "AW / water activity").

Other approaches based on supervised learning methods consist of selecting relevant expansions. In (Xu and Huang, 2007), the authors use SVM approaches (Support Vector Machines) with features based on acronym/expansion information (e.g. length, presence of special characters, context, etc). (Torii et al., 2007) present a comparative study of the main approaches (supervised learning methods, rules-based approaches) by combining domain-knowledge.

Larkey *et al.*'s method (Larkey et al., 2000) uses a search engine to enhance an initial corpus of Web pages useful for acronym detection. To do so, starting from a list of given

acronyms, queries are built and submitted to the AltaVista[1] search engine. Query results are Web pages which URLs are explored, and possibly added to the corpus.

3. Acronym/expansion recognition

Our method of construction of acronym dictionaries is based on two steps detailed in the following subsections.

3.1. Step 1: Extraction of candidates

First, specific punctuation and character markers are taken into account in order to identify acronym/definition pairs (see Figure 1). In this paper, we investigate the extraction of candidates by exploiting the "glosses" of words and paralinguistic markers (i.e. brackets, punctuations, etc.) to detect acronym/definition candidates.

Glosses are spontaneous descriptions identifiable with specific markers (for example, *called*, *i.e.*, and so forth). These ones highlight lexical semantic relationships, e.g. equivalence, specification of the meaning, nomination, hyponomy, hyperonomy.

The abstract pattern of glosses is given by the structure X *marker* $Y_1, Y_2 ... Y_n$ where X and Y_i can be acronyms and/or definitions. The identification and selection of glosses are based on the use of patterns and Web-mining approaches (Mela et al., 2012).

In this paper, we extract candidates based on the gloss markers "(" and ")":

- **Local Pattern 1 [X=acronym, Y_1=definition]:** The first pattern detects Y_1 (definition), between "(" and ")" following the acronym (X). For example, the sentence *"relation empirique entre l'indice de végétation NDVI (Normalized Difference Vegetation Index), mesuré au maximum ..."* allows to extract X = *NDVI* and Y_1 = *Normalized Difference Vegetation Index*.
- **Local Pattern 2 [X=definition, Y_1=acronym]:** The second pattern detects Y_1 (acronym), between "(" and ")" following the definition (X). The beginning of the definition is recognized with the first word of the phrase in upper case. For example, the sentence *" ... la mesure Normalized Difference Vegetation Index (NDVI)"* allows to extract X = *Normalized Difference Vegetation Index* and Y_1 = *NDVI*.

Note that these patterns are independent of languages because the method is based on paralinguistic markers (i.e., brackets in this work). This is very important when languages are mixed, for instance in specialized domains. The example of Figure 1 shows a definition in English (expansion of "NDVI") in an abstract written in French.

In this situation, we are 4 different cases of results:

- **Case 1:** the relevant definition is returned (like previous examples),
- **Case 2:** the extracted phrase contains the relevant definition (i.e. partially relevant, but too large),
- **Case 3:** the extracted phrase is a part of the relevant definition (i.e. partially relevant, but too specific),
- **Case 4:** the extracted phrase is irrelevant.

Both proposed patterns will be evaluated in Section 4. of this paper.

3.2. Step 2: Filtering of candidates

The second step aims at removing irrelevant acronym/definition pairs and deleting irrelevant word(s) from candidate definitions. For this process, we propose to align the acronym letters with the potential definition words, by mapping each acronym letter with the first character of each definition word, respecting the order of words. If the first letter of the candidate definition word can not be aligned with the acronym corresponding character, the following characters (of the word) are taken into account. For instance, this method allows to find that "Extraction Itérative de la Terminologie" is a possible definition of the French acronym EXIT.

4. Evaluation

This paper focuses on the study of a corpus of 2000 paper abstracts provided by Cirad[2]: French research centre working with developing countries to tackle international agricultural and development issues. Table 1 shows that better results are given with the second local pattern. But a lot of cases are partially relevant (i.e. $\sim 40\%$), so we have to improve and enrich this pattern approach.

Patterns	Local pattern 1	Local pattern 2
Number of extracted definitions	78	64
Case 1 (relevant)	31 *(39.7%)*	28 *(43.7%)*
Case 2 (partially relevant)	3 *(3.8%)*	6 *(9.3%)*
Case 3 (partially relevant)	1 *(1.3%)*	18 *(28.1%)*
Case 4 (irrelevant)	43 *(55.1%)*	12 *(18.7%)*

Table 1: Evaluation of extracted definition with patterns.

The evaluation of the acronym/expansion extraction method is conducted on a corpus (general domain) having a reasonable size (7465 words). The experiments based on standard evaluation measures of data-mining domain highlight acceptable results (i.e. Precision: 66.7%, Recall:

[1] www.altavista.com/

[2] http://www.cirad.fr/en/home-page

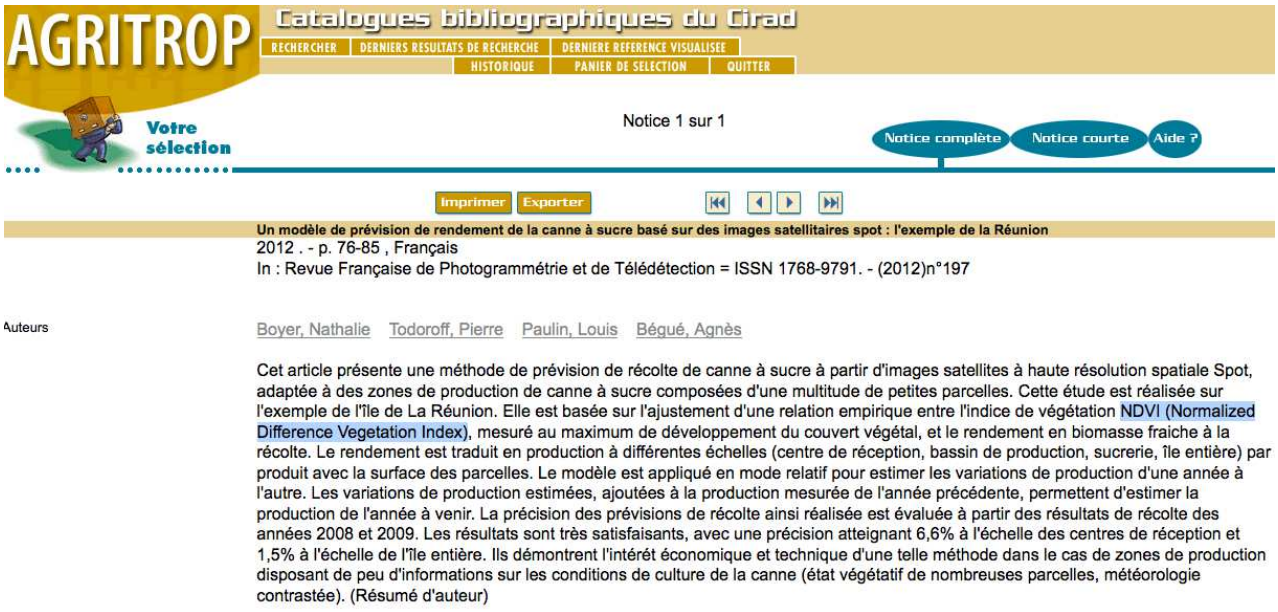

Figure 1: Recognition of the couple *NDVI / Normalized Difference Vegetation Index* in AGRITROP database.

Examples of extracted with Local pattern 1	
NRPS	NonRibosomal Peptide Synthetase
VLE	Virtual Laboratory Environment
BMR	Bois Massif Reconstitué
ATPSM	Agricultural Trade Policy Simulation Model
ASA	Articulation du Semi-aride
CLF	Corynespora Leaf Fall
BASIC	Brésil, Afrique du Sud, Inde, Chine
Examples of extracted with Local pattern 2	
CIAT	Centro international de agricultura tropical
BSV	Banana streak virus
ER	Ehrlichia ruminantium
CSSV	Cacao swollen shoot virus
MAE	Mesures agrienvironnementales
ACMV	African cassava mosaic virus
TYLCV	Tomato yellow leaf curl virus

Table 2: Examples of acronyms/definitions.

80%, F-measure: 72.7%) (Roche and Prince, 2008). We plan to apply the second step of the process (see Section 3.2.) with the pattern approach described in Section 3.1. on the Cirad corpus.
Note that our previous work (Roche and Prince, 2008) uses more global patterns ; then a lot of noise is returned. The pattern approach described in this paper is more specific with better results in term of precision ($\sim$ 40% in this current work vs. 15% in our previous work).

5. Discussion: Towards a Web-mining approach

In this section, we propose to integrate Web-mining measures in order to automatically validate results returned by our approach (Turney, 2001; Mela et al., 2012).

For instance, we can query a search engine with the acronym "BSV" and its possible definition to check on the Web if this association exists. This query should be a disjunction (i.e. OR operator) of the acronym and its possible definition returned with our process (i.e. Banana streak virus). This one returns a larger amount of documents. The conjunction of the acronym and the expansion (i.e. AND operator) enables to return a lower number of documents. But the returned documents are more relevant (i.e. the precision is improved).
In our case, we choose to consider the "hits" given by Google[3] on the examples of Table 2 (i.e. number of pages returned by the search engine based on conjunction). For instance, we have tested the query *"BSV" AND "Banana streak virus"* that returns 7580 pages[4]. All the results (i.e. hits) are given in Table 3. This table shows that hits have generally very high values, this allows us to automatically validate acronym/definition couples. Note that hits of irrelevant couples return lower values (for instance, with the couples "ETM"/"environ 5.000 m3.ha-1", "SIPSA"/"indicateurs, documents, cartes", and so on).

Moreover, we can integrate this kind of information in classical similarity measures, e.g. Dice measure (Smadja et al., 1996). Dice measure can be used to compute a sort of relationship between an acronym (i.e. $acro$) and a definition (i.e. def). In our context, Dice measure (formula (1)) is based on the number of Web pages given by search engines (i.e. hits).

$$Web_{Dice}(acro, def) = \frac{2 \times \text{hits}(acro, def)}{\text{hits}(acro) + \text{hits}(def)} \quad (1)$$

[3] http://www.google.fr/
[4] Queries performed on the 20th of March 2014.

Acronym	Possible definition	Hits (Google)
NRPS	NonRibosomal Peptide Synthetase	230000
VLE	Virtual Laboratory Environment	36900
BMR	Bois Massif Reconstitué	9270
ATPSM	Agricultural Trade Policy Simulation Model	27700
ASA	Articulation du Semi-aride	663
CLF	Corynespora Leaf Fall	22800
BASIC	Brésil, Afrique du Sud, Inde, Chine	21100
CIAT	Centro international de agricultura tropical	75000
BSV	Banana streak virus	7580
ER	Ehrlichia ruminantium	121000
CSSV	Cacao swollen shoot virus	2040
MAE	Mesures agrienvironnementales	951
ACMV	African cassava mosaic virus	90200
TYLCV	Tomato yellow leaf curl virus	354000

Table 3: Examples of acronym/definition and hits scores.

This measure returns the following result with the previous example:

$$Web_{Dice}(BSV, Banana\ streak\ virus)$$

$$= \frac{2 \times \text{hits}(\text{"}BSV\text{"}\ AND\ \text{"}Banana\ streak\ virus\text{"})}{\text{hits}(\text{"}BSV\text{"}) + \text{hits}(\text{"}Banana\ streak\ virus\text{"})}$$

$$= \frac{2 \times 7580}{2840000 + 15400}$$

$$= 0.0053$$

Web_{Dice} can be applied in order to rank couples (see Table 4). This enables to detect relevant acronym/definition pairs (i.e. couples with high Web_{Dice} values).

Acronym	Possible definition	Web_{Dice}
ATPSM	Agricultural Trade Policy Simulation Model	1.3014
TYLCV	Tomato yellow leaf curl virus	0.7167
NRPS	NonRibosomal Peptide Synthetase	0.4423
CIAT	Centro international de agricultura tropical	0.1408
ACMV	African cassava mosaic virus	0.0970
CSSV	Cacao swollen shoot virus	0.0245
VLE	Virtual Laboratory Environment	0.0222
CLF	Corynespora Leaf Fall	0.0208
BSV	Banana streak virus	0.0053
BMR	Bois Massif Reconstitu	0.0046
ER	Ehrlichia ruminantium	0.0004
BASIC	Brsil, Afrique du Sud, Inde, Chine	0.0001
ASA	Articulation du Semi-aride	0
MAE	Mesures agrienvironnementales	0

Table 4: Acronym/definition couples ranked with Web_{Dice}.

6. Conclusion

The process described in this paper is based on the use of specific linguistic markers to detect acronyms. In future work we plan to integrate statistical information and Web-mining approaches in order to improve our methods based on linguistic rules.

Our text-mining system allows us to enrich specialized thesaurus (e.g. MeSH[5], Agrovoc[6]). These thesaurus are useful to automatically annotate texts.

Moreover we plan to investigate a contrastive analysis of English/French corpora in order to give a new point of view of the phenomenon of spontaneous descriptions. A first study on aligned English/French texts reveals frequent regularities of glosses in a multilingual context. The alignment enables to improve the multilingual lexical acquisition of new words and their translations.

7. Acknowledgements

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University Montpellier 2 and CNRS. The author thanks DIST (Scientific and Technical Information Service) for the acquisition of Cirad corpus.

8. References

Chang, J., Schtze, H., and Altman, R. (2002). Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9:612–620.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Larkey, L. S., Ogilvie, P., Price, M. A., and Tamilio, B. (2000). Acrophile: An automated acronym extractor and server. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pages 205–214.

Mela, A., Roche, M., and el Amine Bekhtaoui, M. (2012). Lexical knowledge acquisition using spontaneous descriptions in texts. In *Proceedings of Natural Language Processing and Information Systems Conference (NLDB)*, pages 366–371.

Nenadic, G., Spasic, I., and Ananiadou, S. (2003). Terminology-Driven Mining of Biomedical Literature. *Bioinformatics*, 19(8):938–943.

Okazaki, N. and Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–3095.

Roche, M. and Prince, V. (2008). Managing the acronym/expansion identification process for text-mining applications. *Int. J. Software and Informatics*, 2(2):163–179.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Torii, M., Hu, Z., Song, M., Wu, C., and Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*.

Turney, P. (2001). Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Proceedings of ECML, Lecture Notes in Computer Science*, pages 491–502.

Xu, J. and Huang, Y. (2007). Using SVM to extract acronyms from text. *Soft Comput.*, 11(4):369–373.

Yeates, S. (1999). Automatic extraction of acronyms from text. In *New Zealand Computer Science Research Students' Conference*, pages 117–124.

[5]http://www.nlm.nih.gov/mesh/

[6]http://aims.fao.org/standards/agrovoc/about